Querying, Navigating and Visualizing an Online Library Catalog

Aravindan Veerasamy
Scott Hudson
Shamkant Navathe
{veerasam, hudson, sham}@cc.gatech.edu
College of Computing
801, Atlantic Drive
Georgia Institute of Technology
Atlanta, Georgia 30332-0280
Phone: 404-894-8791
Fax: 404-894-9442

Abstract

We describe the design of an User Interface for a ranked output Information Retrieval system that integrates querying, navigation and visualization in a seamless fashion. Highlights of the system include the following:

- Using a visualization scheme, the interface provides visual feedback to the user about how the query words influence the ranking of retrieved documents.
- By simple drag-and-drop operations of objects on the screen, the interface facilitates a naive end-user in constructing complex structured queries and in providing relevance feedback.
- To suit the evolving information needs of the user, the interface supports navigational features such as browsing documents by specific authors and browsing the Table of Contents of publications.
- The interface integrates an online thesaurus which provides words related to the query that can be used by the user to expand the original query.

By providing a rich set of features, the interface coherently supports a wide spectrum of information gathering tactics for different classes of users.

1 User Interface issues for ranked Information Retrieval systems

User interface issues and interaction techniques for information retrieval have in general received much less attention than system issues like document representation and retrieval algorithms. It is our belief that the interface should portray a system that is in line with the user's needs and information seeking strategies as opposed to one that only supports querying. We have attempted a holistic approach to building an interface that integrates querying, browsing and visualization all in one system [VNH95].

We present an interaction technique for relevance feedback that can gracefully handle feedback at multiple levels of granularity – whole documents, document portions, phrases and words. The interaction technique tries to mimic the user's view that some of the information displayed by the system is in line with the query and wishes to see more of it, while some information is definitely not what the user intended and should be avoided in future. The user can classify almost any piece of information on the screen as positive or negative by dragging and dropping the information object into either a positive area or a negative area on the screen.

Current experimental ranked output IR systems tend to automate the whole gamut of query processing with tools like natural language processing of queries to identify syntactic constructs and thesaurus to automatically expand query terms with related terms. While such approaches may be successful in future, we believe that the system can be more effective by letting the user provide high quality input such as user selection of thesaurus terms. Voorhees [Voo94] mentions that automatic expansion of query terms using thesaurus words has not been very successful. Spink [SS92] however mentions in her study of source of search terms of real users with intermediaries, that about 20% of the search terms in the final query were from a thesaurus. In an effort to maintain the quality of thesaurus terms used to expand the query, we involve the user in the process of selecting thesaurus terms. This is done by integrating an online thesaurus from which the user picks related words to expand his/her query.

Also, in order to provide high quality feedback information to the system and reformulate the query during subsequent iterations, the user would be in a better position if he/she understands the system and has some idea of how the search results were computed. Seeing a demo of the current interface, reference librarians at our university (who are probably among those most willing and able to formulate the best possible search), were perplexed with the ranking of query result documents and were quite concerned about dealing with a system whose retrieval mechanism for the ranking process is not known. This must be contrasted against the ease with which they can figure out why a set of documents were retrieved in response to a boolean query. So that there is no confusion, we are not implying that the reference librarians prefer boolean queries over free-form queries. On the contrary, they feel that a majority of the users would be more comfortable with free-form queries than with boolean queries. But at the same time, they seem concerned about not knowing the ranking process "behind the scene". To portray the system as much less of a black box, and to keep the user more informed about how the query result ranking was computed, we use a visualization scheme that shows how the query results are related to the query words.

In order to shape it into a well-rounded IR system, Bates [Bat89] recommends some browsing features that need to be supported. These features include searching the list of references cited by a particular article, searching the list of articles which cite a particular article, browsing the list of articles written by a particular author, browsing all the articles in a particular journal (issue), browsing physically collocated books in an area. Along these lines, we believe that the system should support a rich set of browsing features to enable users with diverse information needs and searching strategies and to help the user through different stages of knowledge acquisition as highlighted by Belkin's notion of ASK [BOB82]. To this end, the interface facilitates browsing articles by authors and browsing the table of contents of journal issues and conference proceedings.

As mentioned above, we address three interface aspects in our system – interaction techniques for relevance feedback which is discussed in section 2, explaining the ranking of documents by means of visualization which is discussed in section 3 and support for browsing in addition to querying which is discussed in section 4.

2 Interactive Construction of Queries and Relevance Feedback

Searching a text database for information is a highly interactive process with the user constantly refining the query after examining the results of previous iteration until he/she is either satisfied with the results or is completely unsuccessful with the process and gives up. In existing information retrieval systems, the interaction proceeds by the user providing feedback on which of the retrieved documents are relevant to his/her information need. The system uses this information to modify the original query resulting in an improved ranking of retrieved documents. It has also been shown by Spink [SS92] that during iterative query reformulation, users tend to expand the query using search terms from various sources such as a thesaurus, previously retrieved documents and user's background knowledge. Expanding the query with terms from such sources can contribute to retrieval of more relevant documents in the next iteration.

Our interface encourages the interaction between the user and the system by providing the user with a simple interaction techniques to let him/her supply relevance feedback at different levels of granularity: whole documents, document portions, phrases and individual words. Almost any information appearing on the screen can be used for feedback. This is achieved by "drag-and-drop"ping of the feedback object into either a "Positive Objects" window colored green or a "Negative Objects" window colored red. This scheme provides a simple abstraction to the user for classifying any type of information without having to worry about what action to take for what type of information. A typical user session along with the response of the interface for every user action is described below using an example (refer to Figure 1).

- The user types in his free form textual query in the query window. In the example shown in figure 1, the query is "ozone depletion and melanoma"
- As every query word is typed in, the system consults an on-line thesaurus and displays words and phrases related to the query word in an adjacent window.
- At any point during the session the user can "drag-and-drop" any of the related words/phrases into the positive and negative windows. Internally the system expands the query by treating the positive words/phrases as synonyms of the corresponding query word. The negative words/phrases are included in the query with a NOT operator. For example, if for a query word "bank", the phrase "financial institution" is classified as positive and "river bed" is classified as negative, the corresponding internal query would be "#SYNONYM(bank #2[1](financial institution)) #NOT(#2(river bed))". The interface facilitates construction of such structured queries by simple "drag-and-drop" operations. In the example in figure 1, one word related to the query word "melanoma, namely, "skin cancer" has been classified as positive. Internally the systems treats the phrase as a synonym of "melanoma".
- After the user types in the query, the system evaluates the query and displays the titles of top-ranked documents in the "Query Results" window.
- The user examines the query result. Double-clicking any title with the mouse will bring up the full document.
- The user can classify any document as being relevant or non-relevant by "drag-and-drop"ping the document into positive and negative windows. In the example in figure 1, the user has classified two documents titled "CFC-free integral skin foams for steering wheels." and "Video comparator system for early detection of cutaneous malignant melanoma" as positive. The document titled "Symposium on chemistry of the Atmosphere" has been classified as negative.

[1]#2() is the proximity operator in INQUERY specifying that the words inside braces should appear within a distance of 2 of each other in the document.

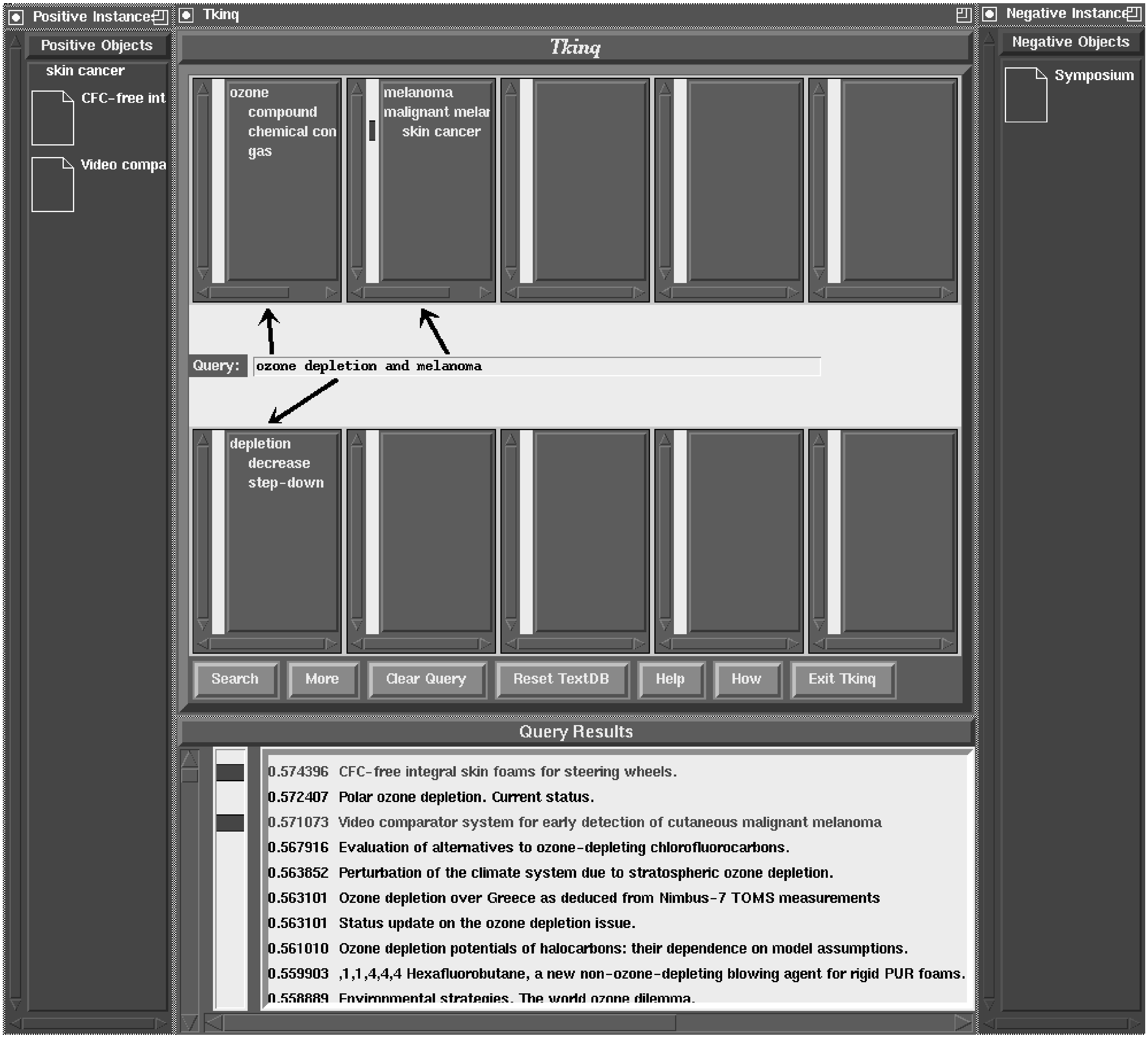

Figure 1: Sample querying session. The window titled "Positive Objects" is colored green and the window titled "Negative Objects" is colored red. When a document is classified as positive/negative, the title of that document in the "Query Results" window is also colored green/red.

- The user can also highlight a portion of a document and "drag-and-drop" that portion into the positive and negative windows. The words in the highlighted document portion are used to expand the query in the next iteration.

- During the next iteration, the reformulated query with the feedback information is processed by the system resulting in an improved ranking of documents.

The positive and negative windows for feedback are aimed at mimicking the user's view that some information is in line with the information need and some not. After an object has been classified as positive (or negative), the system always colors the object green (or red) whenever the object is displayed, thereby reinforcing the user with the fact that the object is being used for relevance feedback. While arguing for the use of direct manipulation techniques for Information Retrieval, Mitev [Mit89] mentions that

> "Parts of document(s), individual word(s), sentences or groups of word(s) displayed could be used directly as something to be input for another search. This could be done, for example, by pointing and 'picking' them on the screen and carrying them across another area of the screen. The user would not have to input them again."

This is precisely what has been accomplished in our interface. In their retrieval system, Campbell [CS] uses a cut-and-paste mechanism for relevance feedback by letting the user add portions of retrieved documents back into the query window.

This section dealt with the interaction technique to let the user provide relevance feedback information to the system. The next section deals with visual feedback from the system on how the query results were computed.

3 Visualization of query results

While systems with a boolean retrieval model retrieve an unordered set of documents in response to a query, ranked output information retrieval systems retrieve a ranked set of documents. While the reason for retrieving a document is fairly clear in the case of a boolean system, the reason why a document is assigned a specific rank is not apparent in the case of a ranked output system. Without knowing how the system computed the ranking of documents, the user will have to treat the retrieval mechanism as a black box. The system stands to gain a lot by keeping the user more informed about the retrieval process of the system. If the user has more information about how the ranking was computed, he/she will be in a better position to reformulate the query for the next iteration. He/she can take into account the deficiencies of the system in adjusting his/her query. It will also help in reinforcing the right mental model.

In our interface, we keep the user informed about the retrieval mechanism by providing visual feedback about how the query results are related to the query words. This is done by a visualization scheme as shown in the figure 2. The visualization reveals the extent to which each query word was responsible for retrieving the set of documents. The visualization consists of a set of histograms, one for every query word (except stop words) typed in by the user, and one histogram for the total query (labeled "Total sum"). All the histograms are placed one below the other with the "Total sum" histogram appearing at the bottom and the query-word-histograms appearing in the order in which query words were typed in. Each histogram consists of a set of vertical bars, one bar for each retrieved document. For the top ranked document, a vertical bar is drawn in the leftmost position (i.e, lowest X coordinate position) in the "Total sum" histogram. The height of the bar is proportional to the weight of the document. (Note that each document is given a weight. The higher the document weight, the more likely it is to be relevant to the query.) For the same document, vertical bars in the same X-coordinate position are also drawn in the query-word-histograms. The height of the vertical bar in any given query-word-histogram is proportional to the weight of the query word in that document. It represents the contribution of the query word in retrieving that document. If the query word does not appear in the document, thereby getting a weight of zero, a bar of zero height is drawn which shows up as an empty space in that X-coordinate position. The second ranked document occupies the next higher X-coordinate to the right and so on upto a maximum of top 200 documents.

The visualization shown in Figure 2 corresponds to the base query with no feedback information from the user. We can see that only fifteen of the top 200 documents have anything to do with melanoma. Almost all of the 200 documents were retrieved because they contained the query words "ozone" and "depletion". Further, of those fifteen documents, only one discusses ozone (the top ranked document – leftmost bar in Figure 2.) Thus we can see clearly that there are not many documents that discuss the the base query about the effects of ozone layer depletion on melanoma. Either there are not many documents about the effects of ozone layer on melanoma or the concept of ozone layer drowns out melanoma. The visualization scenario after providing feedback (that "skin cancer" is a synonym of "melanoma") and computing the results is shown is shown in Figure 3. The figure shows that there are four documents dealing with melanoma and ozone. (Note that the documents which deal with melanoma and it's synonym skin cancer are displayed in the same histogram titled

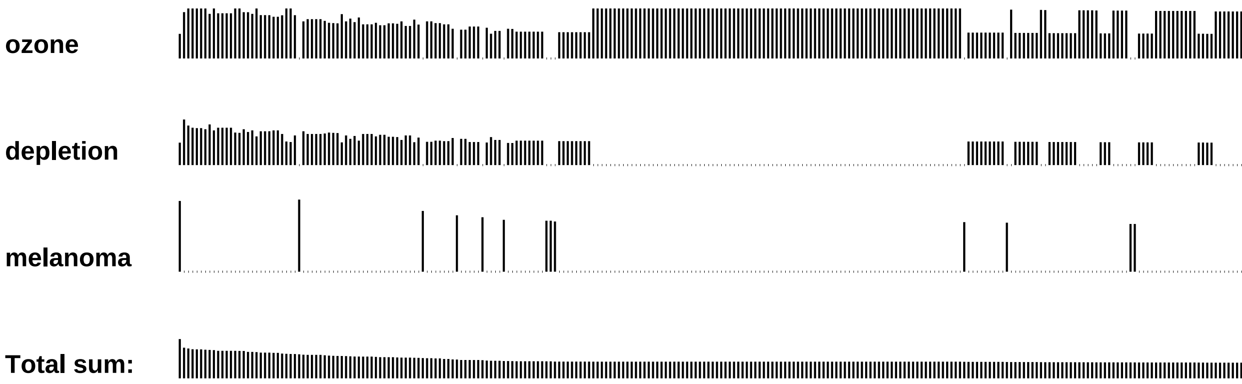

Figure 2: Visualization of results for the base query.

"melanoma", since melanoma and skin cancer represent the same query concept). Thus there are three additional documents retrieved due to the effect of classifying the phrase "skin cancer" as a synonym of "melanoma". But still there are not many documents about melanoma compared to ozone depletion. Our experience with this visualization scheme has shown it to be a useful tool for identifying different facets of the query, as in this case, the facets are melanoma and ozone. The visualization also illustrates which of the query words play a dominant part in retrieving the results and the user has a better idea of what type of query modification is necessary during the next iteration.

4 Browsing

The motivation to integrate browsing features in a querying system has been strongly influenced by [Bat89], [Hil89] and [BOB82]. While it has been argued by all of them that browsing is a central information seeking strategy commonly employed by users, we do not know of any existing online library catalog that integrates browsing and querying. Some examples of browsing activity performed by researchers are:

- While coming across a special issue of a journal or a conference publication devoted to the researcher's are of interest, he/she browses through the table of contents and some articles in the publication.
- On identifying a journal specific to the researcher's are of interest, he/she would want to browse through the publication to keep up-to-date on the developments in the field [Bat89].

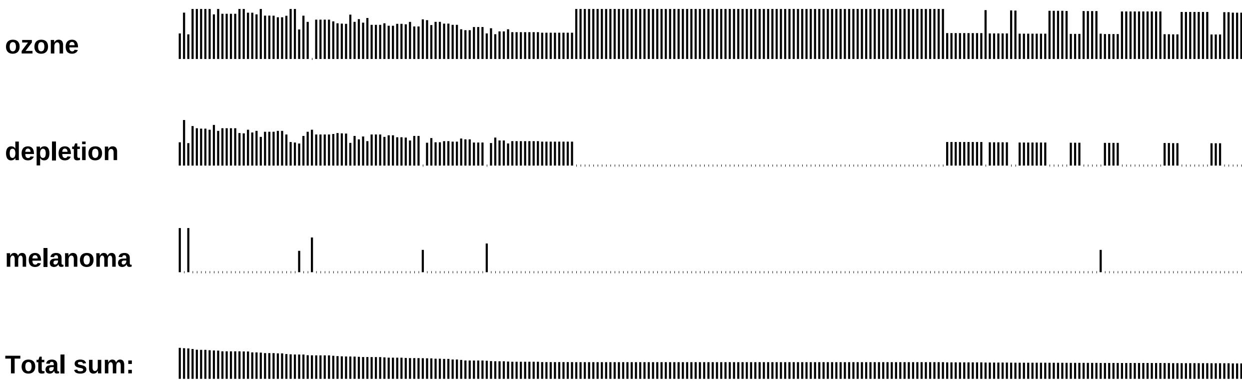

Figure 3: Visualization of results for query with feedback information.

- On discovering that a particular author is working in the same or closely related area, one might want to browse through articles written by that author [Bat89].

In all these cases, the user need not necessarily have a specific information need in mind and may not be able to formulate a query. Conversely, the user might want to browse through the documents as mentioned above while perusing the results of a previously constructed query. While it is true that a user can get the articles written by an author with an "author search" or the articles in a journal with a "journal search", the user may not necessarily want to abandon the results of the current search to initiate a new search author or journal search. The context switch need to initiate a new search can be distracting and disorienting. Browsing can accomplish the "author search" or "journal search" while maintaining the context. Browsing in our system is illustrated in figures 4 and 5. As shown in figure 4, a document display consists of the title of the documents, its authors, pages, name of the journal/conference proceedings, volume, number, part of the publication and the abstract. The portion of the document which can be clicked upon to browse through related information is underlined. As shown in Figure 4, the journal name and the authors are underlined, and hence browsable. Double-clicking an author or journal name from the document display would initiate an internal search in the system, but externally appears as navigation to the user. In our example, double-clicking the journal citation "Wave motion VOL 13 NUM 2" displays the table of contents of the that journal issue. The table of contents is shown in Figure 5. Double-clicking any article from the table of contents would display the whole article. Author browsing is supported in a similar way.

5 Related Work

Numerous studies on user interaction with online library access catalog systems with a boolean retrieval model have been conducted [Spi93, SS92, Dal90, Fid91a, Fid91b, Fid91c]. Spink [Spi93] studies the different forms of user feedback during a retrieval session. In her study, Spink [Spi93] mentions that of the total number of feedback actions by the user, 45% were aimed at adjusting the size of the retrieved set of documents, and about 40% were related to relevancy of documents. Fidel [Fid91a, Fid91b, Fid91c] discusses the issue of user interaction by studying the process of search term selection and searching styles in online library access catalogs. Dalrymple [Dal90] looks at the feedback process from a user-centered perspective. Bates [Bat90] describes a boolean retrieval system which integrates an online thesaurus. None of the above studies involve a ranked output system supporting free-form textual queries. All of the above systems deal with boolean systems only. We believe that there is a significant difference in the way users interact with a boolean system and a ranked output system. The reader is referred to [Har92] and [HB92] for a comparative discussion of boolean systems and ranked output systems. While building our interface we have borrowed valuable ideas from the studies mentioned above. In particular, the need to integrate an on-line thesaurus with the search interface in an easy-to-use fashion and a simple interaction scheme to include words from documents into the query have been influenced by the results of above-mentioned studies. We expect that the studies we plan to conduct with library users using our interface will provide important insights into the ways users react to ranked output systems. It is also expected to give us an idea of the set of the most useful features to be supported by a ranked output information retrieval system.

Walker [Wal87, HB92] describes their OKAPI system which is a ranked output retrieval system for library catalogs. Similarly, Fox [FFS+93] describes their MARIAN system which is also a ranked output system for library catalogs based on vector-space model. While OKAPI and MARIAN have facilities for relevance feedback and query expansion using a thesaurus, they largely lack any means of providing system feedback to the user about how the ranking was computed. The interface we have developed integrates relevance feedback information from the user as well as feedback from the system illustrating the relationship between query results and query words.

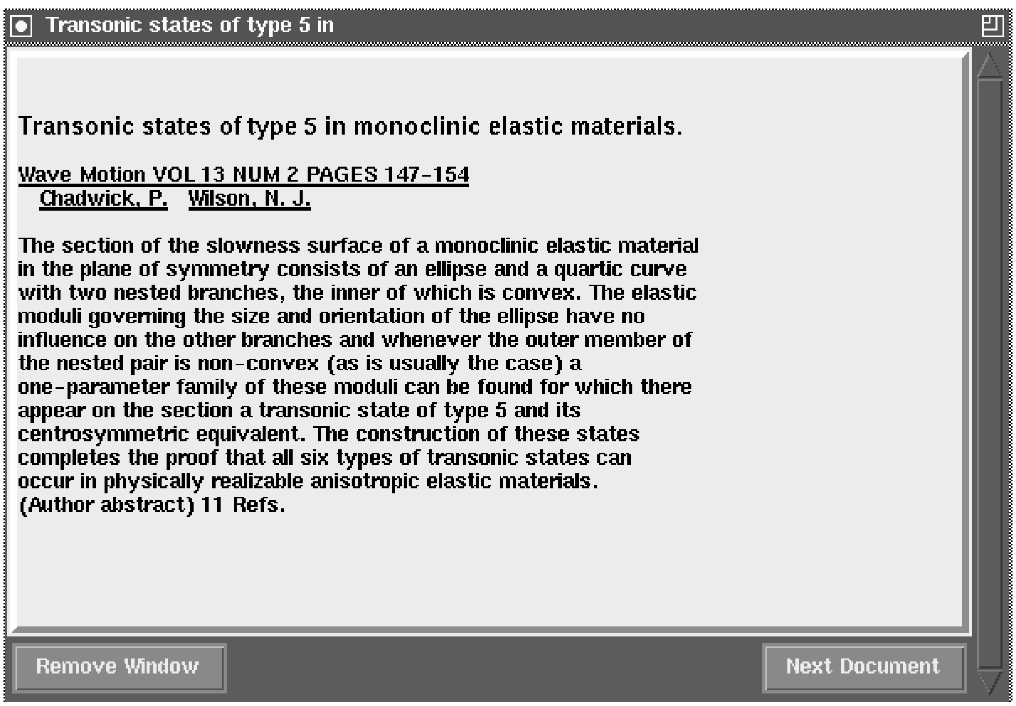

Figure 4: Sample document display. The browsable portions are underlined.

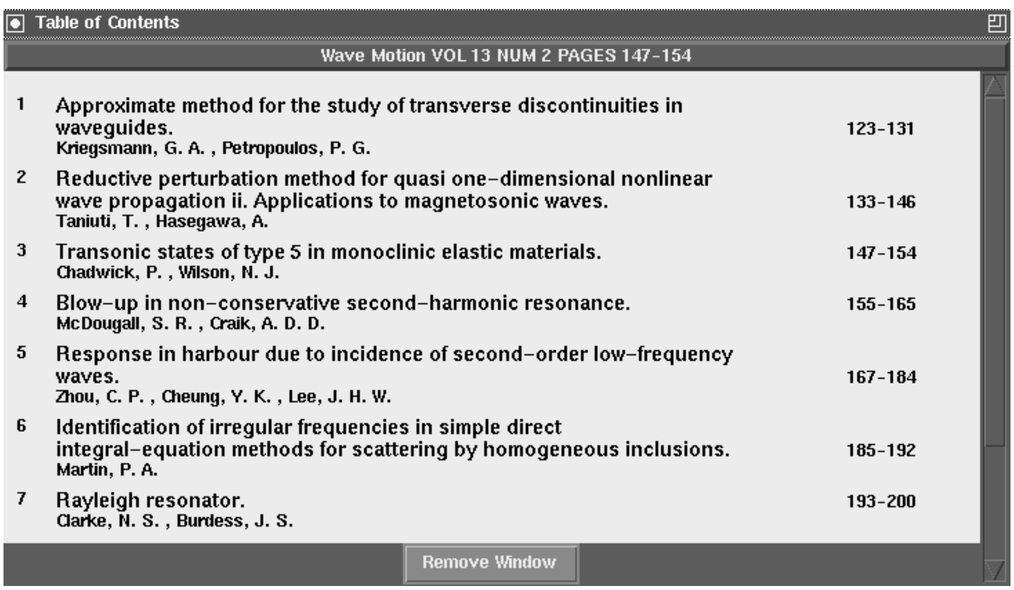

Figure 5: Sample display of table of contents.

A number of visualization schemes for information retrieval systems have also been proposed. The perspective wall [CRM91] describes a visualization scheme which supports browsing of documents. While such a system can not handle qualitative document classifications such as library subject catalogs, it is very useful for visualizing documents based on data which is linear in nature (like date of publication). Other visualization schemes such as [Kor91, Spo94, HKW94] have facilities for viewing a large document space. But visualizing the document space along more than 3 - 4 dimensions simultaneously becomes very cumbersome using their systems. Also, most of them do not support querying with relevance feedback and none of them support query expansion using a thesaurus. The visualization scheme in our interface can gracefully handle much higher number of query word dimensions.

The TileBars work by Marti Heart [Hea95] visually shows the query term ditribution and overlap in retrieved documents. The term distribution in retrieved documents is shown right besides the title of the document. In a number of respects, the reasons and motivations for her work are similar to those of our visualization work [VNH95, VN95]. That both of us seem to have similar motivations behind our work independently of each other reflects on the need for such work. It would be a very interesting exercise to evaluate both TileBars and our visualization work and study their effectiveness in end-user experiments. We are currently undertaking end-user evaluation of our work as part of the interactive track of TREC-4 [TRE95].

Belkin and his group's work [BMC93, BMA+91, HB94] on user interfaces for information retrieval systems elucidates the issues in user interface and interaction techniques for full text retrieval systems. Belkin [BMA+91] mentions that

> ... analysis led to another important conclusion, namely that information systems for end users must support a variety of goals and tasks, but through some common interface or seamless access mechanism to a variety of relevant information sources and system functionalities.

Our interface is a step in that direction by integrating different pieces of information with a visualization scheme and simple interaction techniques.

6 Conclusion & Future work

A prototype interface written in Tcl/Tk [Ous94] using a ranked output information retrieval system, INQUERY [CCH92] for a library catalog, Compendex containing about 300,000 documents has been implemented. The interface facilitates the inherently interactive nature of the information seeking process. "Drag-and-drop" operations form the basis of interaction encouraging the user to provide feedback information to the system and helps in the dialog between the user and the system. Almost any information on the screen can be used by the user to provide feedback information. An online thesaurus, WordNet [MBF+90], is integrated with the interface to form a single system.

The interface also supports a visualization scheme which illustrates how the query results are related to the query words. Visualizing the results of the query keeps the user more informed on how the system computed the ranking of documents. With this information, the user is better equipped to reformulate the query for the next iteration. The interface also has facilities to browse the Table of Contents of publications and to browse the list of articles written by a specific author. It is our opinion that integrating all of the above features in a seamless interface leads to an interplay between different items that is much more beneficial than the sum of the individual items in isolation.

We are in the final stages of implementation, and in future, we intend to test the effectiveness of the interface by conducting studies on how library users interact with the interface and how they react to ranked output systems as opposed to existing boolean systems. We plan to include a domain-specific thesaurus from Compendex and a collection-specific word-association thesaurus if possible.

7 Acknowledgments

We deeply appreciate the help of Dr. Scott Hudson who was instrumental in his guidance at every stage of the interface development. We are thankful to Dr. Bruce Croft for letting us use the INQUERY retrieval system. We are indebted to the Dean of Georgia Tech Library Ms. Miriam Drake and Engineering Information Inc without whom it would have been impossible to use Compendex data for the experiment. Many thanks to Dr. Marti Hearst whose Tcl/Tk code for SMART was a jumping board to build our system. Support from the ARPA contract No. F33615-93-1-1338 is appreciated.

References

[Bat89] Marcia J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407, 1989.

[Bat90] Marcia J. Bates. Design for a subject search interface and online thesaurus for a very large records management database. In *Proceedings of the 53rd Annual Meeting of the American Society for Information Science*, pages 20–28, 1990.

[BMA+91] N.J. Belkin, P.G. Marchetti, M. Albrecht, L. Fusco, S. Skogvold, H. Stokke, and G. Troina. User interfaces for information systems. *Journal of Information Science*, 17:327–344, 1991.

[BMC93] N.J. Belkin, P.G. Marchetti, and C. Cool. Braque: Design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):325–344, 1993.

[BOB82] N.J. Belkin, R.N. Oddy, and H.M. Brooks. Ask for information retrieval: Parts I and II. *Journal of Documentation*, 38(2,3), 1982.

[CCH92] J.P. Callan, W.B. Croft, and S.M. Harding. The inquery retrieval system. In *Third International Conference on Database and Expert Systems Applications*, September 1992.

[CRM91] S. Card, G. Robertson, and J. Mackinlay. The information visualizer, an information workspace. In *Proceedings of CHI 91 Human Factors in Computer Systems.*, 1991.

[CS] I. Campbell and M. Sanderson. Personal communication. University of Glasgow.

[Dal90] P.W. Dalrymple. Retrieval by reformulation in two library catalogs: toward a cognitive model of searching behaviour. *Journal of the American Society for Information Science*, 41(4):272–281, 1990.

[FFS+93] Edward A. Fox, Robert K. France, Eskinder Sahle, Amjad Daoud, and Ben E. Cline. Development of a modern OPAC: From REVTOLC to MARIAN. In Robert Khorfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of sixteenth ACM SIGIR conference*, pages 248–259. ACM SIGIR, June-July 1993.

[Fid91a] Raya Fidel. Searcher's selection of search keys: I. the selection routine. *Journal of the American Society for Information Science*, 42(7):490–500, 1991.

[Fid91b] Raya Fidel. Searcher's selection of search keys: II. controlled vocabulary or free-text searching. *Journal of the American Society for Information Science*, 42(7):501–514, 1991.

[Fid91c] Raya Fidel. Searcher's selection of search keys: III. searching styles. *Journal of the American Society for Information Science*, 42(7):515–527, 1991.

[Har92] Donna Harman. User-friendly systems instead of user-friendly front-ends. *Journal of American Society for Information Science*, 43(2):164–174, 1992.

[HB92] Micheline Hancock-Beaulieu. User friendliness and human-computer interaction in online library catalogues. *Program*, 26(1):29–37, January 1992.

[HB94] Scott Henninger and Nick Belkin. Tutorial on interface issues and interaction strategies for information retrieval systems. In *Human Factors in Computing Systems CHI 94 Conference Companion*, pages 387–388, 1994.

[Hea95] Marti A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of CHI 95, Denver, Colarado.*, 1995.

[Hil89] Charles R. Hildreth. *The online catalogue.* The library association, 1989.

[HKW94] Matthias Hemmje, Clemens Kunkel, and Alexander Willet. Lyberworld – a visualization user interface supporting full text retrieval. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, pages 249–259, 1994.

[Kor91] Robert Korfhage. To see, or not to see – is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR conference on Research and Development in Information Retrieval*, pages 134–141, 1991.

[MBF+90] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.

[Mit89] Nathalie N. Mitev. Ease of interaction and retrieval in online catalogues: contributions of human-computer interaction research. In Charles R. Hildreth, editor, *The online catalogue*, chapter 8, pages 142–176. Library Association Publishing, London, 1989.

[Ous94] John K. Ousterhout. *Tcl and the Tk Toolkit.* Addison-Wesley, 1994.

[Spi93] Amanda Spink. Interaction with information retrieval systems: Reflections of feedback. In *Proceedings of the Annual Meeting of the American Society for Information Science*, pages 115–121, 1993.

[Spo94] Anslem Spoerri. Infocrystal: A visual tool for information retrieval and management. In *Human Factors in Computing Systems CHI 94 Conference Companion*, pages 11–12, 1994.

[SS92] Amanda Spink and Tefko Saracevic. Sources and use of search terms in online searching. In *Proceedings of the 55th Annual Meeting of the American Society for Information Science*, pages 249–255, 1992.

[TRE95] In D.K. Harman, editor, *The Third Text REtrieval Conference (TREC-3).* NIST Special Publication, March 1995.

[VN95] A. Veerasamy and S. Navathe. Querying, navigating and visualizing a digital library catalog. In *Proceedings of the Second International Conference on the Theory and Practice of Digital Libraries*, 1995.

[VNH95] A. Veerasamy, S. Navathe, and S. Hudson. Visual interface for textual information retrieval systems. In *Proceedings of the Third Conference on Visual Database Systems. IFIP 2.6*, 1995.

[Voo94] Ellen M. Voorhees. On expanding query vectors with lexically related words. In Donna K. Harman, editor, *The second Text REtrieval Conference*, pages 223–231, 1994.

[Wal87] Stephen Walker. Okapi: Evaluating and enhancing an experimental online catalog. *Library Trends*, Spring:631–645, 1987.